Estágio Científico e Tecnológico II - EE016

Relatório

Predição de Séries Temporais Baseada em Redes Neurais Artificiais

Submetido à
Faculdade de Engenharia Elétrica e Computação (FEEC)

Departamento de Engenharia de Computação e Automação Industrial (DCA)
Faculdade de Engenharia Elétrica e de Computação (FEEC)
Universidade Estadual de Campinas (UNICAMP)
CEP 13083-852, Campinas - SP

Aluno: João Pedro de Oliveira Pagnan
Orientador: Prof. Levy Boccato

Campinas, 3 de novembro de 2021

1 Introdução

Durante o primeiro semestre deste projeto de iniciação científica, realizamos o estudo de alguns modelos de predição baseados em redes neurais artificiais, bem como dos fundamentos de sistemas com dinâmica caótica. Estes temas foram apresentados no relatório da disciplina EE015 (seções 3.2 e 2.5, respectivamente). Agora, neste relatório referente à disciplina de EE016, vamos apresentar os cenários utilizados para a análise, a metodologia utilizada, os resultados obtidos e, por fim, as conclusões desta pesquisa sobre o desempenho de redes neurais artificiais na predição de séries temporais originadas por sistemas com dinâmica caótica, além de fornecer alguns detalhes sobre as arquiteturas adicionais que foram incorporadas ao trabalho nesta segunda parte da pesquisa: a *gated recurrent unit* (GRU) [1] e a rede neural com estados de eco (ESN, do inglês *echo state network)* [2].

A seção 2 apresenta os quatro cenários escolhidos para a análise do desempenho das redes neurais, sendo dois destes a tempo discreto e dois a tempo contínuo. No caso, os sistemas a tempo discreto foram o mapa de Hénon [3] e o mapa logístico [4]. Já os cenários a tempo contínuo envolveram o sistema de Lorenz [5] e as equações de Mackey-Glass [6].

Por sua vez, na seção 3, discutiremos os dois modelos previamente citados (GRU e ESN) que foram estudados e implementados, juntamente com as redes neurais apresentadas no relatório parcial, nesta segunda parte da pesquisa.

A seção 4 detalha inicialmente a análise de sensibilidade paramétrica feita para cada modelo, indicando os parâmetros testados e os critérios definidos para o processo de busca em grade [7]. Além disso, a seção 4 também apresentará a metodologia utilizada para definir o número de amostras de entrada de cada modelo preditor (nesse caso, chamado de K), além de indicar qual foi a progressão do erro quadrático médio (EQM) em função do valor de K para cada modelo nos quatro cenários.

Por fim, as seções 5 e 6 mostram os resultados e as conclusões obtidas, respectivamente, encerrando, assim, esta pesquisa de iniciação científica e o relatório para a disciplina Estágio Científico e Tecnológico II.

2 Cenários escolhidos

Antes de falarmos sobre os cenários utilizados na análise, vale a pena recordarmos as características principais de sistemas com dinâmica caótica.

Sistemas caóticos se destacam pois, apesar de serem determinísticos, apresentam dependência sensitiva em relação às condições iniciais (DSCI). Dessa forma, duas trajetórias que partem de posições relativamente próximas no espaço de estados podem evoluir de uma forma totalmente distinta devido às não-linearidades presentes que amplificam as diferenças entre essas condições iniciais [8].

De forma resumida, a dinâmica caótica é marcada pela presença dos seguintes aspectos [9]:

1. Forte sensibilidade com respeito às condições iniciais;
2. A evolução temporal das variáveis de estado (parâmetros de ordem do sistema) é rápida e tem uma aparência errática;
3. Um sinal originado por um sistema caótico tem espectro de potências contínuo e de faixa larga;
4. Há uma produção de informação por parte do sistema;
5. Dão origem a atratores estranhos (estruturas topológicas que ditam a evolução temporal do fluxo de um sistema caótico) [10].

Retomados os pontos principais da dinâmica caótica, daremos continuidade à discussão apresentando os cenários escolhidos para a análise. Vale mencionar que, na simulação numérica dos quatro sistemas foram geradas 5000 amostras para cada série temporal. Além disso, nos sistemas multidimensionais, como o mapa de Hénon e o sistema de Lorenz, consideramos apenas a variável de estado x na previsão.

2.1 Sistema de Lorenz

O sistema de Lorenz foi um dos sistemas dinâmicos caóticos a tempo contínuo abordados nessa pesquisa. Este sistema foi um dos primeiros grandes trabalhos envolvendo a noção de regime caótico, sendo considerado por muitos a pesquisa que inaugurou a área [11].

Através de simulações numéricas de dinâmicas atmosféricas, o matemático e meteorologista Edward Norton Lorenz observou uma dependência sensitiva às condições iniciais em certos sistemas dinâmicos [8]. Dando continuidade a seus experimentos, Lorenz modelou, através de três equações diferenciais, o fluxo de um fluido em um volume uniformemente aquecido na camada inferior e uniformemente resfriado na camada superior [5], as quais são mostradas a seguir:

$$\frac{dx}{dt} = -\sigma \cdot (x - y) \tag{1a}$$

$$\frac{dy}{dt} = x \cdot (\rho - z) - y \tag{1b}$$

$$\frac{dz}{dt} = x \cdot y - \beta \cdot z \tag{1c}$$

sendo σ, ρ e β constantes reais, estando relacionadas a certas características físicas do sistema, como o número de Prandtl, o número de Rayleigh e as dimensões do volume que o fluido ocupa [8].

Utilizando $\sigma = 10$, $\rho = 28$ e $\beta = 8/3$, Lorenz demonstrou que esse sistema de equações diferenciais exibe comportamento caótico, sendo que a maioria das condições iniciais $[x(0)\ y(0)\ z(0)]^T$ leva à convergência para um atrator estranho (nesse caso, atrator de Lorenz).

A figura 1 indica a série temporal em x, que foi utilizada em nossa análise, para $[x(0)\ y(0)\ z(0)]^T = [0.1\ 0\ 0]^T$, e o atrator de Lorenz para a trajetória. Para a simulação, os parâmetros do sistema foram configurados com os mesmos valores utilizados por Lorenz (exibidos no parágrafo anterior), e foi utilizado $dt = 0.01$ para resolver as equações diferenciais numericamente.

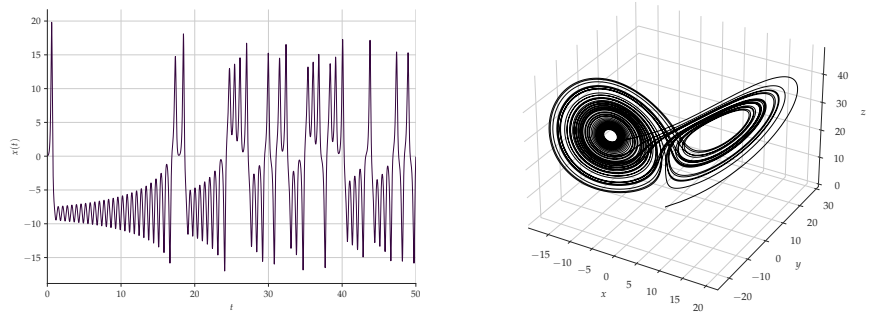


Figura 1: À esquerda, a série temporal em x do sistema de Lorenz simulado e, à direita, o diagrama de fases correspondente à simulação.

2.2 Mapa de Hénon

O mapa de Hénon foi um dos sistemas a tempo discreto escolhidos para esta pesquisa. Esse sistema foi proposto pelo astrônomo e matemático francês Michel Hénon em 1976 como um modelo simplificado de uma seção de Poincaré do atrator de Lorenz, sendo descrito pelas equações abaixo [3]:

$$x[n+1] = y[n] + 1 - a \cdot (x[n])^2 \tag{2a}$$

$$y[n+1] = b \cdot x[n] \tag{2b}$$

onde a e b são valores reais.

A presença de dinâmica caótica neste sistema discreto irá depender dos valores dos parâmetros a e b. Hénon mostrou em sua pesquisa que, para $a = 1.4$ e $b = 0.3$, há a presença de um atrator estranho no diagrama de fases desse sistema dinâmico discreto.

Vale mencionar que o mapa de Hénon realiza um mapeamento de dois pontos, chamados de pontos fixos. Para os valores dos parâmetros a e b mencionados anteriormente, tais pontos são dados por:

$$x = \frac{\sqrt{609} - 7}{28} \approx 0.631354477$$

$$y = \frac{3(\sqrt{609} - 7)}{280} \approx 0.189406343$$

Um desses pontos está sobre o atrator e é instável. Tal instabilidade é confirmada quando é realizada uma análise da trajetória com outros pontos próximos a este, onde percebe-se que, dependendo da região do atrator pela qual o ponto em análise se aproxima do ponto fixo, a trajetória irá divergir ou convergir para o ponto no atrator.

A figura 2 mostra a série temporal referente à variável x e o atrator obtido com a simulação para $[x[0]\ y[0]]^T = [1\ 0]^T$.

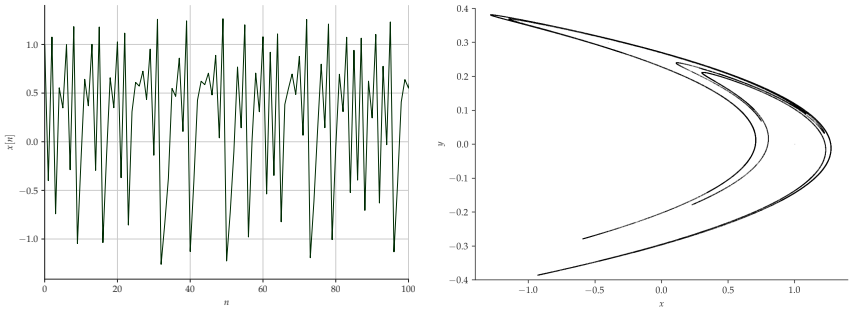


Figura 2: À esquerda, as cem primeiras iterações da série temporal em x do mapa de Hénon e, à direita, o atrator correspondente à simulação.

2.3 Mapa logístico

Descrito em 1976 pelo biólogo e matemático Robert May, o mapa logístico representa uma das formas de modelar a população de uma determinada espécie em certos instantes de tempo [4]. A equação a diferenças que descreve esse sistema pode ser vista abaixo:

$$x[n+1] = r \cdot x[n] \cdot (1 - x[n]) \tag{3}$$

sendo $x[n]$ um número real entre 0 e 1 que representa a razão entre o tamanho atual da população e o tamanho máximo desta, enquanto r é um valor real entre 0 e 4 que representa a taxa de crescimento desta população.

Dependendo do valor de r, o sistema descrito em (3) pode ou não operar em caos. Robert May, através de análises numéricas [11], viu que para valores maiores que 3.44949, o sistema entra em regime caótico. Interessantemente, ao aumentar mais o valor do parâmetro r, o cientista percebeu que há certas zonas de estabilidade, onde o sistema volta a convergir para um ciclo-limite.

Inicialmente, após sair do regime caótico, a dinâmica terá período 2, e, com o aumento de r, o sistema começa a ter período 4, 8, 16, 32, e assim por diante, até voltar, novamente, ao cenário caótico. Esse padrão de ir do caos à estabilidade para, então, voltar ao regime caótico, se repete até que r atinja seu valor máximo.

Quando o físico e matemático Mitchell Feigenbaum estudou esse fenômeno, também presente em outros sistemas caóticos e em certas estruturas fractais, percebeu

que a razão entre o comprimento de dois intervalos sucessivos de bifurcação tende a $\delta \approx 4.66920$ [12]. Esse valor é chamado de constante de Feigenbaum, não estando relacionado a nenhuma outra propriedade matemática ou constante da natureza conhecida até então.

Como o estudo visa analisar o desempenho para sistemas caóticos, foi utilizado $r = 3.86$, que, conforme será visto no diagrama de bifurcação abaixo, faz com que a série temporal dada pela equação (3) opere em caos.

A figura 3 indica a série temporal obtida partindo de $x[0] = 0.5$ e o diagrama de bifurcação, onde a faixa vermelha representa $r = 3.86$, para este sistema de equações.

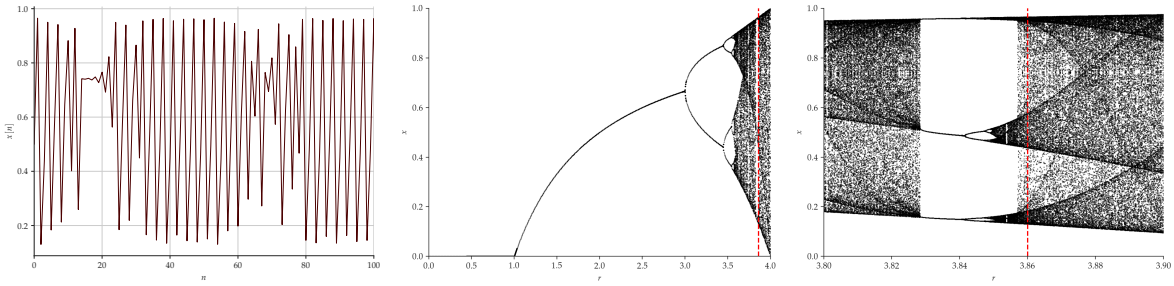


Figura 3: À esquerda, as cem primeiras iterações da série temporal do mapa logístico, ao centro, o diagrama de bifurcação deste sistema e, à direita, a seção do diagrama de bifurcação que contém o valor de r utilizado.

2.4 Equações de Mackey-Glass

Por fim, o último sistema caótico simulado, também contínuo, está associado às equações de Mackey-Glass, as quais modelam o controle hormonal da produção de células brancas do sangue e podem ser vistas abaixo [6]:

$$\frac{dP(t)}{dt} = \frac{\beta_0 \cdot \theta^n}{\theta^n + P(t-\tau)^n} - \gamma \cdot P(t) \tag{4a}$$

$$\frac{dP(t)}{dt} = \frac{\beta_0 \cdot \theta^n \cdot P(t-\tau)}{\theta^n + P(t-\tau)^n} - \gamma \cdot P(t) \tag{4b}$$

sendo $P(t)$ a densidade de tais células em um instante de tempo e β_0, θ, n, τ e γ valores reais relacionados a certos parâmetros hormonais de um organismo, geralmente sendo determinados experimentalmente.

Neste caso, conforme demonstrado em [6], a equação (4b) exibe comportamento caótico para valores mais altos de τ. Além disso, os pesquisadores perceberam que regimes caóticos dessa equação estão correlacionados com certos problemas fisiológicos nos nossos organismos [6].

Para a simulação numérica, utilizamos $n = 10$, $\gamma = 0.1$, $\beta = 0.2$, $\theta = 1$, $\tau = 22$, $dt = 1.0$ e $P(0^-) = 0.1$, gerando novamente 5000 amostras. A série e o atrator obtidos podem ser vistos na figura 4.

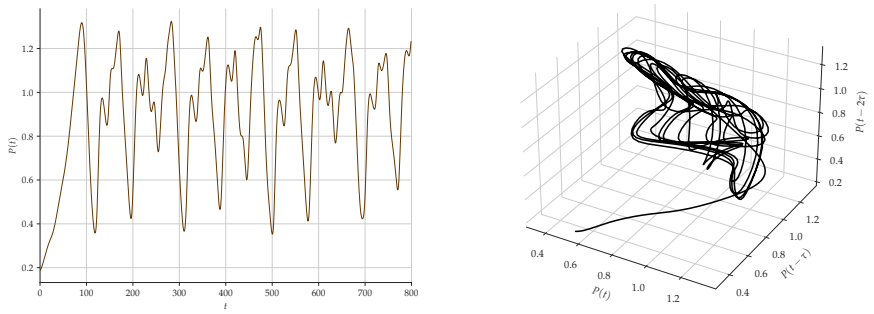


Figura 4: À esquerda, a série temporal da equação (4b) exibida de $t = 0$ a $t = 800$ e, à direita, o atrator correspondente à simulação.

Com isso, finalizamos a exposição sobre os cenários de sistemas com dinâmica caótica escolhidos para a análise comparativa. Na próxima seção, falaremos a respeito dos dois tipos de redes neurais que incorporamos à pesquisa nesta segunda etapa. Em seguida, apresentaremos a metodologia aplicada no treinamento e análise de todos os modelos de previsão estudados, assim como as particularidades para cada caso.

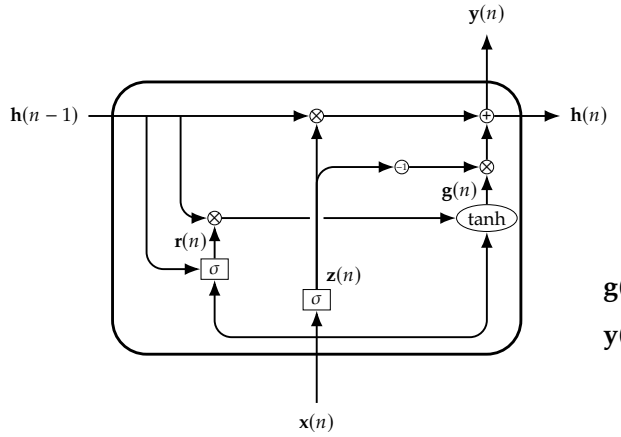
3 Modelos estudados

Como o estudo de redes neurais artificiais foi iniciado na primeira metade da pesquisa, uma exposição mais detalhada dos fundamentos destes modelos (no caso, as redes MLP e as rede recorrentes LSTM) pode ser vista no relatório parcial.

Assim, as próximas duas subseções abordarão especificamente as redes recorrentes GRU e ESN.

3.1 *Gated Recurrent Unit* (GRU)

Como já mencionamos anteriormente, as redes recorrentes *Gated Recurrent Unit* possuem células computacionais semelhantes às células das redes LSTM, apresentadas no relatório parcial. Sua estrutura interna pode ser vista na figura 5, juntamente com as equações que a descrevem.



$$\mathbf{z}(n) = \sigma\left(\mathbf{W}_z[\mathbf{h}(n-1), \mathbf{x}(n)] + \mathbf{b}_z\right)$$
$$\mathbf{r}(n) = \sigma\left(\mathbf{W}_r[\mathbf{h}(n-1), \mathbf{x}(n)] + \mathbf{b}_r\right)$$
$$\mathbf{g}(n) = \tanh\left(\mathbf{W}_g[\mathbf{r}(n) \otimes \mathbf{h}(n-1), \mathbf{x}(n)] + \mathbf{b}_g\right)$$
$$\mathbf{y}(n) = \mathbf{h}(n) = \mathbf{z}(n) \otimes \mathbf{h}(n-1) + (1 - \mathbf{z}(n)) \otimes \mathbf{g}(n)$$

Figura 5: Estrutura e equações de uma célula GRU.

Ao contrário da LSTM, a célula GRU explora um único vetor de estados $\mathbf{h}(n)$. No caso, o vetor $\mathbf{z}(n)$ determina se a memória armazenada em $\mathbf{h}(n-1)$ será mantida ou apagada, enquanto o vetor $\mathbf{g}(n)$ traz a informação nova que será agregada ao vetor de estados, sendo um reflexo do estado anterior e da entrada atual ($\mathbf{x}(n)$).

Intuitivamente, pelo fato de as GRUs serem, em essência, uma versão mais simples das redes LSTM, não seria surpreendente se as LSTM alcançassem um melhor desempenho. Entretanto, conforme verificado em [1], a GRU pode ser superior à LSTM em determinadas aplicações. Quanto à implementação em si, existe uma grande similaridade entre GRU e LSTM (especialmente se trabalhamos com pacotes que fornecem estas camadas para uso dentro de uma arquitetura neural, como é o caso do Keras/Tensorflow).

3.2 *Echo State Network* (ESN)

As redes neurais com estados de eco também são modelos recorrentes para processamento da informação, à semelhança da LSTM e da GRU. No entanto, apresentam um modo de operação e um esquema de treinamento bem diferentes [13].

A figura 6 apresenta a estrutura interna de uma rede ESN.

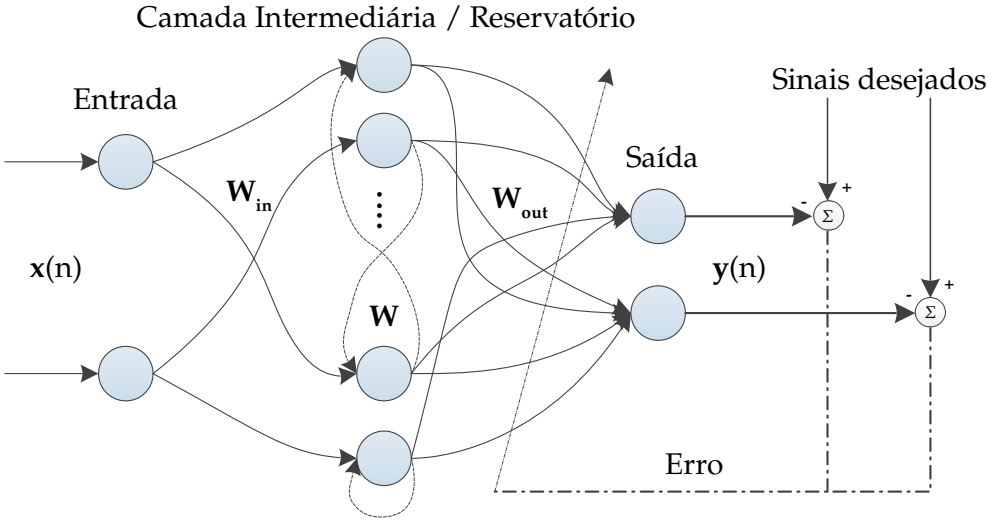


Figura 6: Estrutura típica de uma rede ESN (figura adaptada de [13]).

As equações em (5) descrevem a ESN, sendo $\mathbf{x}(n)$ o vetor de entrada, $\mathbf{W}_{in} \in \mathbb{R}^{N \times K}$ os pesos da camada de entrada, $\mathbf{u}(n)$ o vetor que representa as ativações dos N neurônios não-lineares totalmente conectados do reservatório, representando os estados da rede, $\mathbf{W} \in \mathbb{R}^{N \times N}$ os pesos das conexões recorrentes do reservatório, $\mathbf{f}(\cdot)$ denota as funções de ativação das unidades internas da rede, $\mathbf{W}_{out} \in \mathbb{R}^{L \times N}$ é a matriz dos pesos da camada de saída e, por fim, $\mathbf{y}(n)$ denota as saídas da rede.

$$\mathbf{u}(n+1) = \mathbf{f}(\mathbf{W}_{in} \cdot \mathbf{x}(n+1) + \mathbf{W} \cdot \mathbf{u}(n)) \tag{5a}$$

$$\mathbf{y}(n+1) = \mathbf{W}_{out} \cdot \mathbf{u}(n+1) \tag{5b}$$

O grande diferencial prático desta rede é que os pesos das conexões recorrentes dentro do reservatório $\mathbf{W}$ são ajustados com valores fixos antes do treinamento da camada de saída. Em suma, esses parâmetros são determinados tendo em vista a propriedade de estados de eco, a qual assegura que o vetor de estados da rede tende a ser guiado por um histórico recente das entradas da rede, não mais sofrendo influência dos estados iniciais, desde que condições específicas ligadas à matriz de pesos recorrentes ($\mathbf{W}$) sejam satisfeitas. Conforme demonstrado em [14] e [15], existem alguns métodos simples para a criação aleatória dos pesos que garantam a propriedade.

A definição formal das propriedades de estado de eco é feita sob as seguintes condições [13]:

1. os sinais de entrada são extraídos de um espaço compacto [1] $\mathcal{U}$;
2. os estados da rede estão sempre contidos em um conjunto compacto $\mathcal{A} \subset \mathbb{R}^N$ de estados admissíveis, o que significa que a operação de atualização do vetor $\mathbf{u}(n)$, mostrada na equação (5a), sempre preserva os estados dentro de $\mathcal{A}$.

[1] A definição de espaços compactos foi apresentada em [16].

Utilizando essas restrições, o autor da arquitetura ESN apresentou duas novas condições para a existência das propriedades de estado de eco.

A primeira condição mostra que, caso o máximo valor singular da matriz de pesos do reservatório **W**, em módulo, esteja dentro do círculo unitário ($|\sigma_{max}(\mathbf{W})| < 1$), a rede recorrente apresenta estados de eco [17, 18].

A segunda condição estabelece a não-existência de estados de eco em função do raio espectral, i.e., do autovalor de maior módulo da matriz de pesos internos **W**, denotado por $\rho_s(\mathbf{W})$: se $\rho_s(\mathbf{W}) > 1$, então a rede não possui estados de eco [17]. Para isto, além das restrições apresentadas anteriormente, é assumido que a entrada é nula.

Neste caso, considerando que os pesos internos da rede **W** representam uma espécie de reservatório de dinâmicas que, por sua vez, deve gerar um conjunto de comportamentos dinâmicos o mais diversificado possível, a matriz de pesos **W** deve ser criada com um certo grau de esparsidade, conforme demonstrado em [17]. Dessa forma, um conjunto de comportamentos dinâmicos é obtido sem um conhecimento prévio do sinal que se deseja modelar.

Por fim, vale mencionar que o ajuste dos parâmetros na camada de saída pode ser realizado com um procedimento de regressão linear (neste caso, utilizando amostras passadas do sinal a ser previsto), sem necessitar de um treinamento iterativo, como nas outras redes neurais estudadas nesta pesquisa. Esse fato economiza tempo de processamento e evita os conhecidos problemas de desvanecimento e de explosão do vetor gradiente, assim como outras instabilidades que podem ocorrer no treinamento de uma rede recorrente.

Com isso, finalizamos o restante da exposição teórica desta pesquisa. A próxima seção descreverá a metodologia utilizada nesta análise.

4 Metodologia

Nesta segunda parte da pesquisa de iniciação científica foram implementados modelos preditores com base nas quatro redes neurais artificiais estudadas (MLP, LSTM, GRU e ESN) para os quatro cenários exibidos na seção 2.

Iniciaremos a discussão com a apresentação da busca em grade realizada para todos os modelos, de forma a determinar os parâmetros ótimos das redes neurais em cada cenário. Em seguida, utilizando os melhores parâmetros, avaliamos a progressão do erro quadrático médio em função do número de amostras de entrada K do modelo preditor. Por fim, comparamos qual foi a média e o desvio padrão do EQM com o melhor valor de K de cada modelo nos quatro cenários. Em todos os experimentos, o horizonte de predição utilizado foi $L = 3$. Assim, iremos predizer o valor da terceira iteração à frente do valor atual da série temporal.

Vale reforçar que, para todos os modelos e cenários, utilizamos 85% das amostras das séries temporais para treinamento (ou, equivalentemente, os primeiros 4250 valo-

res). Além disso, o conjunto de validação, utilizado na busca em grade para avaliar a predição de uma configuração testada, corresponde a 10% dos dados de treinamento (as últimas 425 amostras). Assim, os dados de treinamento efetivamente correspondem às primeiras 3825 amostras, os dados de validação correspondem às 425 amostras depois destas e, por fim, o conjunto de teste corresponde aos últimos 750 instantes de tempo. Essa proporção entre os dados de treinamento, de teste e de validação foi mantida para todas as redes analisadas.

Em seguida, na seção 4.2, será apresentada a análise da autocorrelação parcial das séries temporais, a qual norteou a definição das faixas de valores de K para cada cenário. Por fim, na seção 4.3, mostramos os testes realizados com cada valor de K e com cada modelo de previsão, de forma a encontrar a melhor configuração para esse parâmetro.

4.1 Configurações utilizadas

Conforme dito anteriormente, foi realizada uma busca em grade com validação cruzada para determinar as melhores configurações de cada rede neural para a predição das séries temporais em cada cenário. Para cada arquitetura, um conjunto de valores candidatos foi gerado para cada hiperparâmetro e todas as combinações possíveis foram avaliadas tendo em vista o desempenho do modelo em dados de validação.

Houve algumas diferenças entre o processo realizado para a rede MLP e para as redes recorrentes tendo em vista o caráter destes modelos. Como a relação temporal entre as amostras não é importante para a MLP, a busca em grade considerou uma validação cruzada com 4 *folds* nos dados de treinamento. Ou seja, para cada configuração, foi avaliado o desempenho desta rede na predição de quatro subconjuntos das quatro séries temporais escolhidas.

Já nas redes recorrentes, foi considerado um esquema de validação do tipo *holdout*. Esse processo dividiu o conjunto de treinamento (novamente composto por 85% dos dados gerados) em 4 seções. Cada seção era composta por uma fração dos dados de treinamento, sendo que cada seção incluía a seção anterior no seu conjunto de dados. Por exemplo, a segunda seção obtida pelo *holdout* inclui a primeira seção e mais algumas amostras, além de uma subdivisão de validação que será utilizada para avaliar o resultado. Com isso, obtém-se conjuntos sequenciais de dados para avaliação do desempenho. Esse procedimento é necessário para as redes recorrentes, pois a relação temporal entre os dados de entrada deve ser levada em conta.

Por fim, durante esta etapa de busca dos hiperparâmetros, consideramos que $K = 4$.

Com esses detalhes expostos, iniciaremos a exposição com a busca em grade realizada na rede MLP para, em seguida, falarmos sobre as configurações obtidas para as redes recorrentes.

4.1.1 MLP

Na busca em grade para a MLP, avaliamos diferentes tamanhos de *batch* (2, 4, 8, 16, 32), se haveria ou não uma camada de *batch normalization* na entrada da rede MLP, diferentes opções de função de ativação dos neurônios na camada intermediária (SELU, ReLU, ELU, tanh, Sigmoid), esquemas alternativos para a inicialização dos pesos (Glorot, He, Lecun, tanto com distribuição normal, quanto uniforme), diferentes quantidades de neurônios na camada intermediária (5, 10, 15, 20, 30, 50, 75, 100) e, também, diferentes valores para a taxa de aprendizagem (0.001, 0.003, 0.005, 0.008, 0.01). Em todos os testes, consideramos uma rede MLP com uma única camada intermediária.

Ao final do processo completo de busca, foram identificadas as configurações "ótimas"da rede MLP em cada cenário, as quais são exibidas na tabela 1.

Cenário	*Batch normalization*	Tamanho do *batch*	Ativação	Inicialização	Nº de neurônios	Taxa de aprendizagem
Mapa de Hénon	Não	8	Sigmoid	Glorot Normal	50	0.003
Mapa logístico	Não	2	tanh	Glorot Uniforme	10	0.003
Sistema de Lorenz	Não	2	SELU	Lecun Normal	50	0.001
Equações de Mackey-Glass	Não	4	tanh	Glorot Normal	5	0.001

Tabela 1: Melhores parâmetros para a rede MLP nos cenários em análise.

4.1.2 LSTM e GRU

Para as redes recorrentes LSTM e GRU, foi avaliado o tamanho do *batch* (2, 4, 8, 16, 32), a inicialização dos pesos (Glorot Uniforme, Glorot Normal), o número de neurônios recorrentes na camada intermediária (5, 10, 15, 20, 30, 50, 75, 100), e a taxa de aprendizagem (0.001, 0.003, 0.005, 0.008, 0.01), novamente utilizando apenas uma camada intermediária e mantendo a função de ativação usual da célula recorrente (tanh). Por conta disso, para o sistema de Lorenz, foi feito um ajuste de escala para evitar a saturação dos neurônios.

As tabelas 2 e 3 exibem as melhores configurações obtidas em cada cenário para as redes LSTM e GRU, respectivamente.

Cenário	Tamanho do *batch*	Inicialização	Nº de neurônios	Taxa de aprendizagem
Mapa de Hénon	4	Glorot Normal	15	0.005
Mapa logístico	2	Glorot Uniforme	100	0.008
Sistema de Lorenz	4	Glorot Uniforme	15	0.003
Equações de Mackey-Glass	2	Glorot Uniforme	50	0.003

Tabela 2: Melhores parâmetros para a rede LSTM nos cenários em análise.

Cenário	Tamanho do *batch*	Inicialização	Nº de neurônios	Taxa de aprendizagem
Mapa de Hénon	4	Glorot Normal	30	0.003
Mapa logístico	2	Glorot Normal	100	0.003
Sistema de Lorenz	8	Glorot Uniforme	30	0.001
Equações de Mackey-Glass	2	Glorot Uniforme	10	0.005

Tabela 3: Melhores parâmetros para a rede GRU nos cenários em análise.

4.1.3 ESN

Para esta rede foi realizada uma busca em grade para determinar o número de neurônios no reservatório (30, 50, 70, 90, 100, 120, 140, 160, 180, 200, 240, 280, 320, 360, 400, 440, 480, 500) e o valor do raio espectral (0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 0.95, 0.96, 0.97, 0.98, 0.99), que determina o maior autovalor da matriz de pesos do reservatório. Por tratar-se de uma rede recorrente, também foi utilizado o processo de *holdout* descrito na seção 4.1. Vale também mencionar que a taxa de vazamento, que representa a velocidade com a qual o reservatório atualiza suas dinâmicas, foi fixada em 0.9 para todos os cenários.

A tabela 4 exibe as melhores configurações para a ESN em cada um dos cenários. Um ponto interessante a destacar é que os melhores valores do raio espectral são relativamente pequenos e estão bem afastados de 1.0, o que significa que o reservatório tende a desenvolver uma memória mais curta e, assim, esquecer mais rapidamente o passado. Tal característica parece fazer sentido considerando o caráter caótico das séries temporais exploradas no trabalho.

Parâmetro	Mapa de Hénon	Mapa logístico	Sistema de Lorenz	Equações de Mackey-Glass
Nº de neurônios	500	500	120	500
Raio espectral	0.1	0.1	0.2	0.4

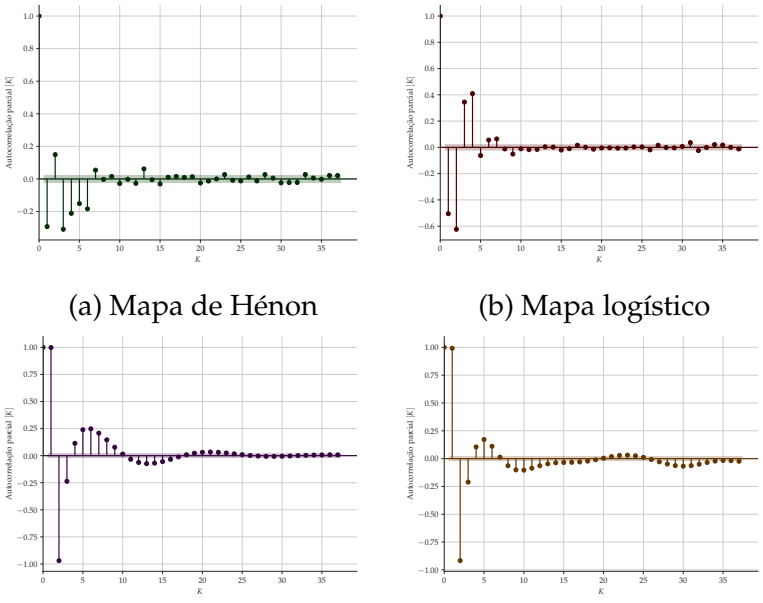
Tabela 4: Melhores parâmetros para a rede ESN nos cenários em análise.

Com isso, concluímos a apresentação das melhores configurações para as redes neurais avaliadas em todos os cenários escolhidos. Na próxima seção, falaremos sobre outra análise realizada, dessa vez não sobre os modelos em teste, mas sim sobre certas características das séries temporais utilizadas.

4.2 Estudo da autocorrelação parcial das séries temporais

Com as melhores configurações para cada rede e cenário obtidas, foi analisada a progressão do erro quadrático médio para cada valor de K. A faixa de valores para K a ser testada foi determinada utilizando a autocorrelação parcial das séries temporais de cada sistema analisado, levando em considerações os atrasos (ou afastamentos temporais entre amostras) que apresentam correlações mais significativas.

A figura 7 mostra a autocorrelação parcial para as séries temporais utilizadas. É interessante perceber a rápida diminuição da correlação conforme o atraso entre amostras aumenta, especialmente nos casos dos sistemas discretos.



(a) Mapa de Hénon (b) Mapa logístico

(c) Sistema de Lorenz (d) Equações de Mackey-Glass

Figura 7: Autocorrelação parcial para as séries temporais utilizadas.

Vale mencionar que a autocorrelação parcial encontra a correlação dos resíduos (assim denominados porque são obtidos após a remoção dos efeitos já explicados pelos primeiros atrasos) com o próximo valor de atraso, removendo correlações entre este e atrasos passados. Logo, essa medida descreve somente a relação direta entre duas observações da série afastadas por K instantes de tempo.

A tabela 5 mostra a faixa de valores para K a ser testada obtida com a análise descrita nos parágrafos anteriores.

	Mapa de Hénon	**Mapa logístico**	**Sistema de Lorenz**	**Equações de Mackey-Glass**
Faixa de valores	$1 \sim 9$	$1 \sim 8$	$1 \sim 10$	$1 \sim 7$

Tabela 5: Faixa de valores inteiros para K a ser testada em cada cenário para todas as redes.

4.3 Análise do melhor valor para K

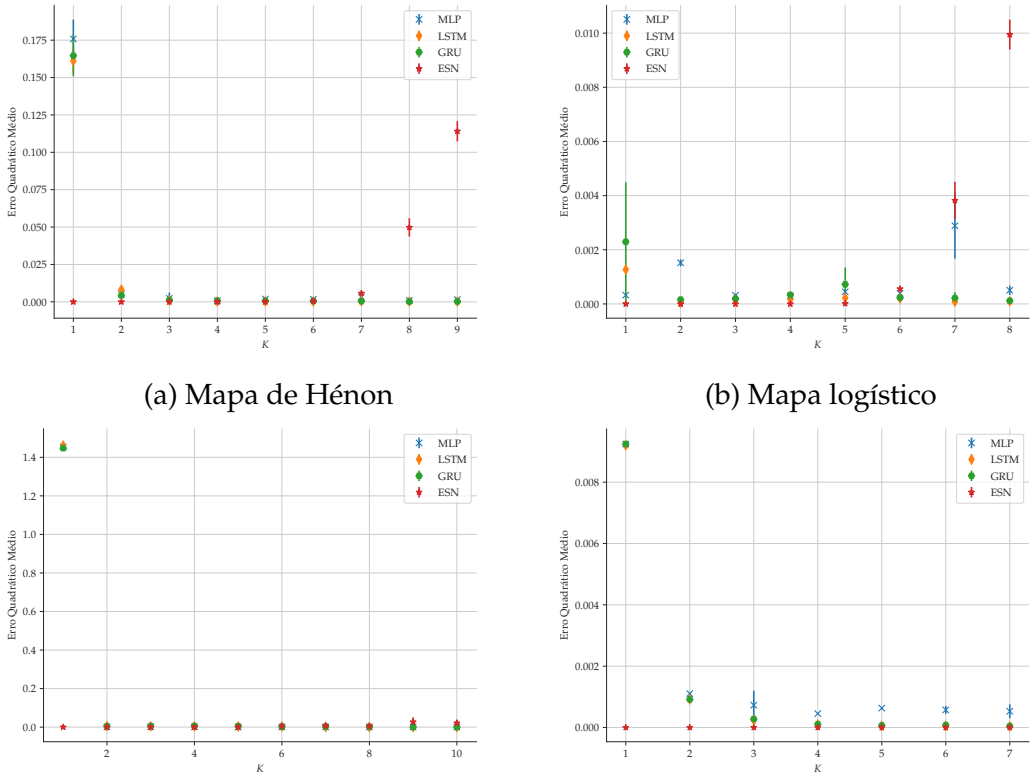
Com as faixas de valores de K estabelecidas, avaliamos a progressão do erro quadrático médio para diferentes configurações desse parâmetro.

Para realizar esse procedimento, cada rede (com as configurações ótimas) foi treinada utilizando 85% dos dados gerados, sendo que 10% dos dados de treinamento foram utilizados como o conjunto de validação (nas redes MLP, LSTM e GRU). Em seguida, com o modelo treinado, foi avaliado o EQM no conjunto de teste (que corresponde às últimas 750 amostras). Esse processo foi realizado 5 vezes para cada K, obtendo assim um valor médio e desvio-padrão para cada modelo e em cada cenário.

Além disso, como o ajuste de parâmetros da ESN possui solução em forma fechada, não foi necessário utilizar um conjunto de validação no treinamento. No caso da MLP, LSTM e GRU, o processo iterativo de ajuste dos pesos utilizou um conjunto de validação de forma a seguir o procedimento de *early stopping* para evitar o sobreajuste da rede.

A figura 8 mostra a progressão do EQM observada em cada cenário para cada uma das redes estudadas.

Nas tabelas 6a, 6b, 6c e 6d, constam os valores médios e desvios-padrão referentes ao EQM para cada valor K dos modelos avaliados, no cenário do mapa de Hénon, do mapa logístico, do sistema de Lorenz, e das equações de Mackey-Glass, respectivamente.



(a) Mapa de Hénon

(b) Mapa logístico

(c) Sistema de Lorenz

(d) Equações de Mackey-Glass

Figura 8: Progressão do erro quadrático médio para cada valor de K nos cenários avaliados para todas as redes neurais testadas.

K	MLP	LSTM	GRU	ESN
1	$(1.758 \pm 0.130) \times 10^{-1}$	$(1.612 \pm 0.081) \times 10^{-1}$	$(1.646 \pm 0.139) \times 10^{-1}$	$(4.657 \pm 3.054) \times 10^{-9}$
2	$(5.190 \pm 1.620) \times 10^{-3}$	$(8.223 \pm 1.426) \times 10^{-3}$	$(4.315 \pm 1.958) \times 10^{-3}$	$(9.764 \pm 3.727) \times 10^{-9}$
3	$(2.346 \pm 3.765) \times 10^{-3}$	$(9.028 \pm 7.506) \times 10^{-4}$	$(1.252 \pm 1.082) \times 10^{-3}$	$(8.699 \pm 2.645) \times 10^{-8}$
4	$(6.415 \pm 4.726) \times 10^{-4}$	$(2.498 \pm 0.854) \times 10^{-4}$	$(7.557 \pm 2.607) \times 10^{-4}$	$(1.595 \pm 0.425) \times 10^{-6}$
5	$(1.892 \pm 0.214) \times 10^{-3}$	$(2.256 \pm 1.282) \times 10^{-4}$	$(6.827 \pm 3.646) \times 10^{-4}$	$(3.355 \pm 0.519) \times 10^{-5}$
6	$(1.739 \pm 1.251) \times 10^{-3}$	$(2.341 \pm 0.471) \times 10^{-4}$	$(5.686 \pm 2.254) \times 10^{-4}$	$(5.113 \pm 1.617) \times 10^{-4}$
7	$(3.239 \pm 1.034) \times 10^{-3}$	$(4.771 \pm 2.439) \times 10^{-4}$	$(4.486 \pm 0.877) \times 10^{-4}$	$(5.731 \pm 0.882) \times 10^{-3}$
8	$(9.436 \pm 2.631) \times 10^{-4}$	$(1.578 \pm 0.689) \times 10^{-4}$	$(2.411 \pm 1.391) \times 10^{-4}$	$(4.978 \pm 0.620) \times 10^{-2}$
9	$(1.514 \pm 0.805) \times 10^{-3}$	$(2.151 \pm 0.605) \times 10^{-4}$	$(3.007 \pm 2.156) \times 10^{-4}$	$(1.141 \pm 0.068) \times 10^{-2}$

(a) Mapa de Hénon

K	MLP	LSTM	GRU	ESN
1	$(3.275 \pm 0.349) \times 10^{-4}$	$(1.269 \pm 0.144) \times 10^{-3}$	$(2.296 \pm 2.195) \times 10^{-3}$	$(1.082 \pm 1.165) \times 10^{-6}$
2	$(1.519 \pm 0.131) \times 10^{-3}$	$(4.836 \pm 4.925) \times 10^{-5}$	$(1.589 \pm 0.424) \times 10^{-4}$	$(1.316 \pm 2.068) \times 10^{-8}$
3	$(3.250 \pm 0.495) \times 10^{-4}$	$(1.470 \pm 0.352) \times 10^{-4}$	$(1.965 \pm 0.849) \times 10^{-4}$	$(1.381 \pm 0.436) \times 10^{-8}$
4	$(2.943 \pm 0.462) \times 10^{-4}$	$(1.683 \pm 0.378) \times 10^{-4}$	$(3.322 \pm 1.297) \times 10^{-4}$	$(5.314 \pm 1.738) \times 10^{-7}$
5	$(4.488 \pm 0.631) \times 10^{-4}$	$(2.193 \pm 0.832) \times 10^{-4}$	$(7.226 \pm 6.237) \times 10^{-4}$	$(1.402 \pm 0.298) \times 10^{-5}$
6	$(3.918 \pm 0.475) \times 10^{-3}$	$(2.073 \pm 1.147) \times 10^{-4}$	$(2.312 \pm 1.397) \times 10^{-4}$	$(5.552 \pm 0.952) \times 10^{-4}$
7	$(2.887 \pm 1.216) \times 10^{-3}$	$(8.216 \pm 5.138) \times 10^{-5}$	$(2.159 \pm 2.128) \times 10^{-4}$	$(3.821 \pm 0.688) \times 10^{-3}$
8	$(5.066 \pm 1.723) \times 10^{-4}$	$(9.880 \pm 5.323) \times 10^{-5}$	$(1.173 \pm 0.678) \times 10^{-4}$	$(9.947 \pm 0.553) \times 10^{-3}$

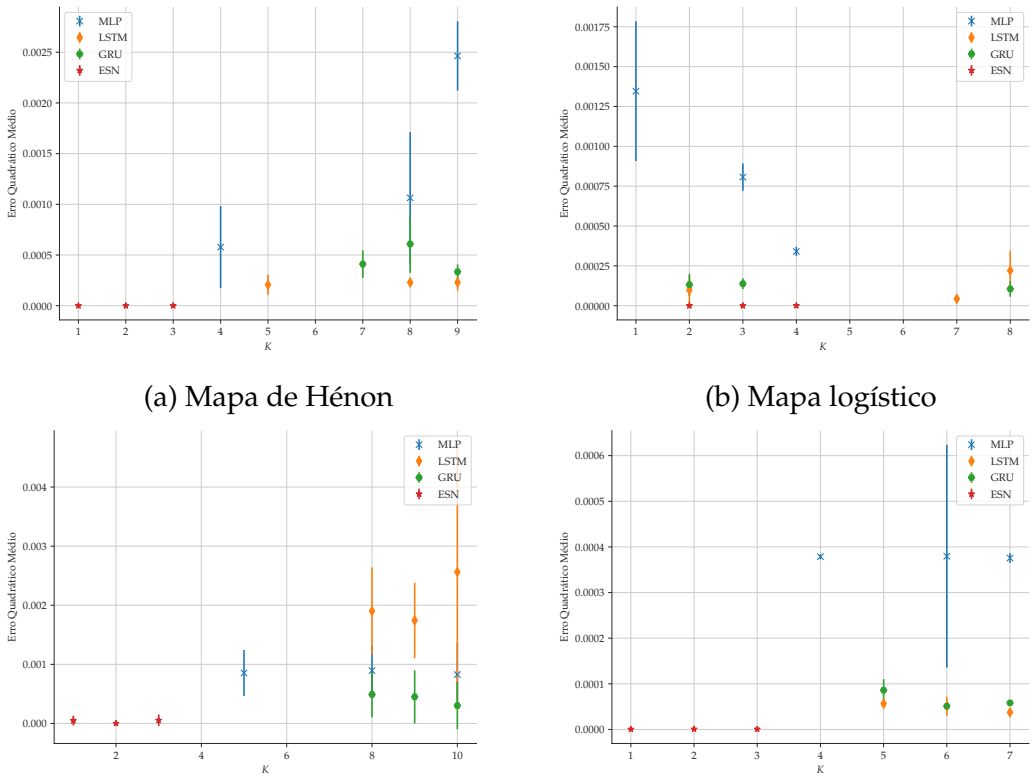
(b) Mapa logístico

K	MLP	LSTM	GRU	ESN
1	$(1.453 \pm 0.005) \times 10^{0}$	$(1.461 \pm 0.012) \times 10^{0}$	$(1.447 \pm 0.004) \times 10^{0}$	$(1.216 \pm 3.706) \times 10^{-5}$
2	$(8.919 \pm 1.357) \times 10^{-4}$	$(5.251 \pm 2.017) \times 10^{-3}$	$(4.897 \pm 0.767) \times 10^{-3}$	$(2.305 \pm 4.065) \times 10^{-6}$
3	$(1.568 \pm 0.476) \times 10^{-3}$	$(3.934 \pm 1.895) \times 10^{-3}$	$(5.308 \pm 1.334) \times 10^{-3}$	$(4.007 \pm 4.187) \times 10^{-5}$
4	$(1.336 \pm 0.360) \times 10^{-3}$	$(3.830 \pm 1.041) \times 10^{-3}$	$(5.503 \pm 1.162) \times 10^{-3}$	$(1.708 \pm 1.799) \times 10^{-4}$
5	$(5.770 \pm 1.322) \times 10^{-4}$	$(3.019 \pm 0.907) \times 10^{-3}$	$(4.433 \pm 1.670) \times 10^{-3}$	$(6.577 \pm 5.963) \times 10^{-4}$
6	$(8.002 \pm 1.270) \times 10^{-4}$	$(2.409 \pm 1.427) \times 10^{-3}$	$(2.318 \pm 0.871) \times 10^{-3}$	$(4.674 \pm 5.859) \times 10^{-3}$

Após avaliarmos o desempenho dos modelos para as faixas de K obtidas com a análise da autocorrelação parcial das séries temporais, foram selecionados os três melhores valores para esse parâmetro de cada modelo, ou seja, os que tiveram as três menores médias de erro quadrático médio, para uma etapa de retreinamento e reavaliação.

Novamente, utilizando os três melhores valores de K, realizamos o processo mencionado no início desta seção, cinco vezes para cada K.

A figura 9 mostra a nova progressão obtida em cada cenário para os modelos avaliados, e as tabelas 7a, 7b, 7c e 7d, exibem os valores numéricos das médias e desvios-padrão para os valores selecionados do parâmetro K.



(a) Mapa de Hénon

(b) Mapa logístico

(c) Sistema de Lorenz

(d) Equações de Mackey-Glass

Figura 9: Progressão do erro quadrático médio para os três melhores valores de K nos cenários avaliados para todas as redes neurais testadas.

Finalmente, depois de visualizarmos a progressão do EQM nos três melhores valores de K para cada caso, selecionamos os melhores para o último teste realizado, novamente, seguindo o procedimento descrito no início desta seção.

A próxima seção exibe os resultados obtidos, juntamente com certos trechos das predições onde a diferença entre os modelos foi perceptível.

Modelo	K	Erro Quadrático Médio
MLP	4	$(5.780 \pm 4.047) \times 10^{-4}$
	8	$(1.064 \pm 0.648) \times 10^{-3}$
	9	$(2.464 \pm 0.341) \times 10^{-3}$
LSTM	8	$(2.287 \pm 0.332) \times 10^{-4}$
	9	$(2.293 \pm 0.813) \times 10^{-4}$
	5	$(2.063 \pm 0.994) \times 10^{-4}$
GRU	8	$(6.083 \pm 2.839) \times 10^{-4}$
	9	$(3.389 \pm 0.727) \times 10^{-4}$
	7	$(4.104 \pm 1.354) \times 10^{-4}$
ESN	1	$(3.251 \pm 0.536) \times 10^{-9}$
	2	$(1.311 \pm 0.366) \times 10^{-8}$
	3	$(8.867 \pm 2.472) \times 10^{-8}$

(a) Mapa de Hénon

Modelo	K	Erro Quadrático Médio
MLP	4	$(3.403 \pm 0.292) \times 10^{-4}$
	3	$(8.069 \pm 0.859) \times 10^{-4}$
	1	$(1.346 \pm 0.438) \times 10^{-3}$
LSTM	2	$(9.874 \pm 8.623) \times 10^{-5}$
	7	$(4.248 \pm 2.074) \times 10^{-5}$
	8	$(2.206 \pm 1.250) \times 10^{-4}$
GRU	8	$(1.054 \pm 0.489) \times 10^{-4}$
	2	$(1.320 \pm 0.687) \times 10^{-4}$
	3	$(1.379 \pm 0.335) \times 10^{-4}$
ESN	2	$(9.254 \pm 2.344) \times 10^{-9}$
	3	$(1.157 \pm 0.500) \times 10^{-8}$
	4	$(6.064 \pm 1.851) \times 10^{-7}$

(b) Mapa logístico

Modelo	K	Erro Quadrático Médio
MLP	5	$(8.542 \pm 3.905) \times 10^{-4}$
	10	$(8.266 \pm 5.422) \times 10^{-4}$
	8	$(8.945 \pm 4.267) \times 10^{-4}$
LSTM	10	$(2.564 \pm 2.149) \times 10^{-3}$
	8	$(1.903 \pm 0.738) \times 10^{-3}$
	9	$(1.744 \pm 0.638) \times 10^{-4}$
GRU	10	$(3.020 \pm 3.998) \times 10^{-4}$
	9	$(4.512 \pm 4.481) \times 10^{-4}$
	8	$(4.905 \pm 3.868) \times 10^{-4}$
ESN	2	$(1.792 \pm 1.593) \times 10^{-6}$
	1	$(4.833 \pm 8.416) \times 10^{-5}$
	3	$(5.384 \pm 9.854) \times 10^{-5}$

(c) Sistema de Lorenz

Modelo	K	Erro Quadrático Médio
MLP	4	$(3.786 \pm 0.054) \times 10^{-4}$
	7	$(3.756 \pm 0.105) \times 10^{-4}$
	6	$(3.795 \pm 2.441) \times 10^{-3}$
LSTM	7	$(3.750 \pm 1.016) \times 10^{-5}$
	6	$(5.089 \pm 2.077) \times 10^{-5}$
	5	$(5.638 \pm 0.208) \times 10^{-5}$
GRU	7	$(5.817 \pm 0.633) \times 10^{-5}$
	5	$(8.594 \pm 2.413) \times 10^{-5}$
	6	$(5.101 \pm 0.938) \times 10^{-5}$
ESN	1	$(4.812 \pm 0.378) \times 10^{-7}$
	2	$(6.493 \pm 0.572) \times 10^{-7}$
	3	$(7.789 \pm 0.221) \times 10^{-7}$

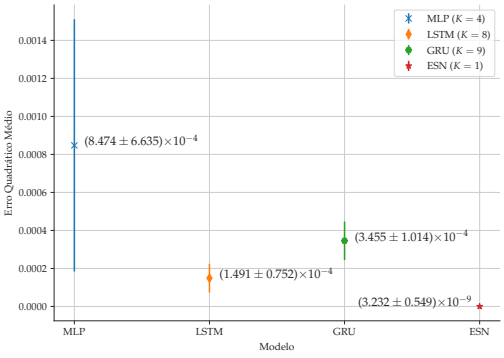
(d) Equações de Mackey-Glass

Tabela 7: Valores médios e desvios-padrão do erro quadrático médio obtidos para cada K nos cenários utilizadas.

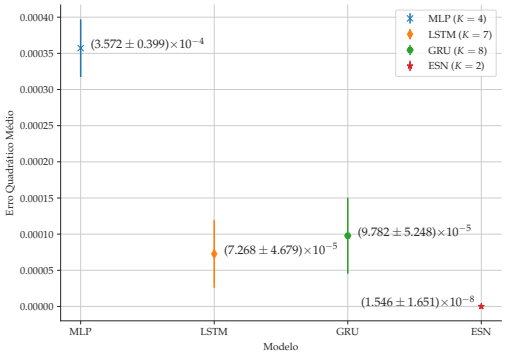
5 Resultados

Após identificarmos o melhor valor de K para cada modelo, realizou-se novamente o processo mencionado na seção anterior e foi obtido o EQM de teste para as melhores configurações (parâmetros e K) para cada modelo, em todos os cenários.

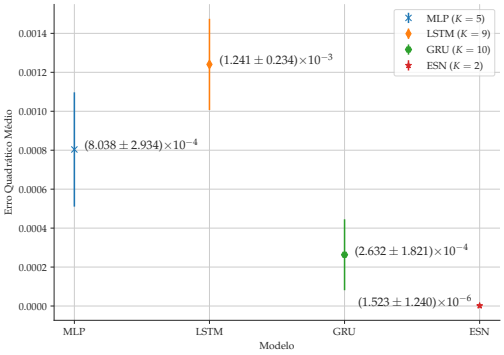
A figura 10 exibe o comparativo dos melhores desempenhos das redes neurais analisadas, e a figura 11 mostra uma comparação da predição em si de cada modelo em certos trechos das séries temporais nos quais a diferença foi mais perceptível visualmente.



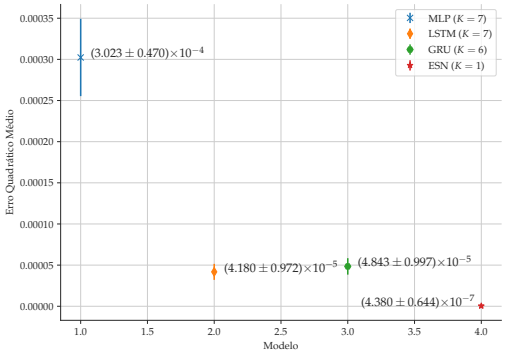
(a) Mapa de Hénon



(b) Mapa logístico

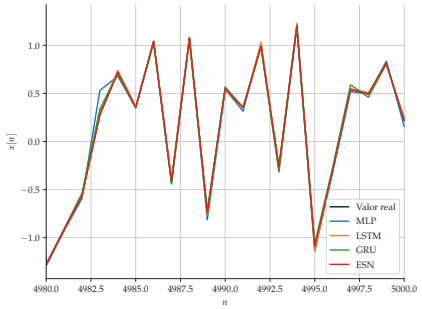


(c) Sistema de Lorenz

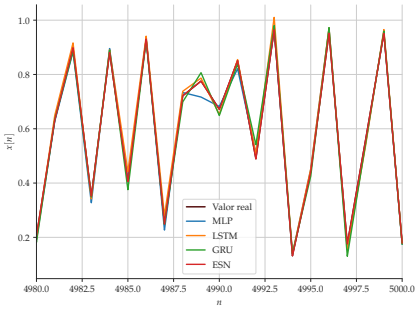


(d) Equações de Mackey-Glass

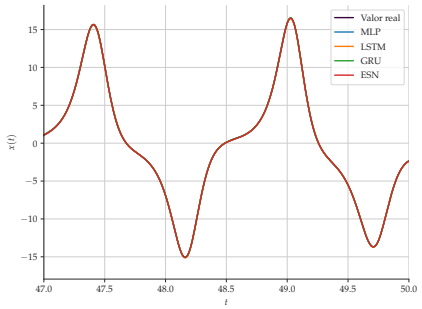
Figura 10: Comparação do melhor desempenho obtido por cada modelo nos cenários testados.



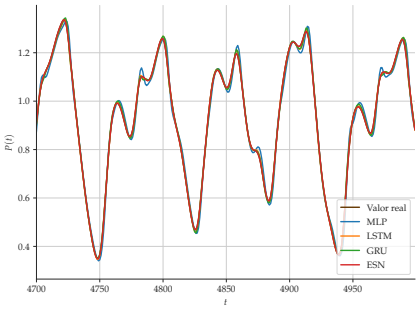
(a) Mapa de Hénon



(b) Mapa logístico



(c) Sistema de Lorenz



(d) Equações de Mackey-Glass

Figura 11: Comparação da predição realizada por cada modelo para os quatro sistemas caóticos.

6 Análise e Conclusão

Analisando os resultados obtidos, percebe-se que, com exceção do cenário do sistema de Lorenz, a rede MLP foi consideravelmente pior do que as redes recorrentes. Também percebe-se que, dentre as redes recorrentes, a ESN obteve um EQM bem inferior ao obtido pela LSTM e pela GRU.

O desempenho inferior da rede MLP com relação às redes recorrentes provavelmente decorre do fato de que a relação temporal presente na LSTM, GRU e ESN auxilia na modelagem da dinâmica da série temporal. Já o pior desempenho da LSTM na série temporal do sistema de Lorenz provavelmente está relacionado aos efeitos estocásticos presentes no ajuste dos pesos sinápticos dessa rede neural que, conforme indicado em [19], é uma dificuldade em seu treinamento.

Algo interessante de mencionar é que, no geral, os desempenhos de todos os modelos estudados foram consideravelmente bons nas séries temporais dos sistemas a tempo contínuo (sistema de Lorenz e equações de Mackey-Glass), sendo que as diferenças foram mais pronunciadas na série temporal do mapa de Hénon e do mapa logístico. Provavelmente, a maior suavidade presente nas séries de Lorenz e de Mackey-Glass facilita a modelagem do preditor, além do fato de que as séries temporais mencionadas são bem menos erráticas se comparadas às séries de Hénon e do mapa logístico.

O principal resultado observado foi que o modelo preditor utilizando a rede com estados de eco obteve um desempenho bem superior aos outros modelos, em todos os cenários. O erro quadrático médio foi tão baixo que, observando a predição nos dados de teste, praticamente não há diferença entre os valores reais e os valores previstos. Essa superioridade da ESN também é realçada na imagem comparativa do EQM, que indica que a ESN atingiu patamares de erro cerca de 100 ou até 10000 vezes menores que os outros modelos.

Esse resultado, aliado ao fato de que o treinamento da ESN é bem menos custoso computacionalmente se comparado ao das outras redes, mostra que esta rede é uma boa alternativa para futuros estudos de modelos preditores de séries temporais. Além disso, conforme mostrado em outros trabalhos como [20, 2, 13], a ESN também é uma boa ferramenta para outras tarefas de extração de informação, como em equalização de canais e separação de fontes.

Como sugestão de trabalhos futuros, pode ser avaliada a eficácia da ESN em reconstruir atratores através de séries temporais experimentais de sistemas caóticos. Se o desempenho para essa tarefa for tão bom quanto o obtido na predição das séries estudadas, a ESN pode tornar-se uma ferramenta ainda mais poderosa para a modelagem de sistemas com dinâmica caótica.

Com isso, finalizamos esta pesquisa de iniciação científica. Através dos processos descritos neste relatório final e, anteriormente, no relatório parcial, o aluno teve contato com ferramentas modernas de predição de séries temporais, úteis não só para séries originadas por dinâmicas caóticas, como também para cenários relacionados a outros tipos de sistemas.

Além disso, o aluno teve uma exposição a outras ferramentas de aprendizado de máquina que não necessariamente estão relacionadas com problemas de predição, mas que podem ser úteis para outros tipos de projetos de pesquisa que o estudante possa participar no futuro.

Logo, esse projeto forneceu os princípios para que o aluno possa atuar junto a diversas áreas do conhecimento, como a matemática, a física e até mesmo a biologia, e desenvolver pesquisas científicas bem fundamentadas, cumprindo, assim, o principal objetivo de um projeto de iniciação científica.

Referências

[1] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[2] H. Jaeger, "Echo state network," *scholarpedia*, vol. 2, no. 9, p. 2330, 2007.

[3] M. Hénon, "A two-dimensional mapping with a strange attractor," *Communications in Mathematical Physics*, vol. 50, pp. 69–77, feb 1976.

[4] R. M. May, "Simple mathematical models with very complicated dynamics," *Nature*, vol. 261, pp. 459–467, jun 1976.

[5] E. N. Lorenz, "Deterministic nonperiodic flow," *Journal of atmospheric sciences*, vol. 20, no. 2, pp. 130–141, 1963.

[6] M. C. Mackey and L. Glass, "Oscillation and chaos in physiological control systems," *Science*, vol. 197, no. 4300, pp. 287–289, 1977.

[7] A. Géron, *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems*. O'Reilly Media, 2019.

[8] N. Fiedler-Ferrara and C. P. C. do Prado, *Caos: uma introdução*. Editora Blucher, 1994.

[9] R. R. de Faissol Attux, "Sobre dinâmica caótica e convergência em algoritmos de equalização autodidata," dissertação (mestrado), Universidade Estadual de Campinas, Faculdade de Engenharia Elétrica e de Computação, Campinas, SP, 2001.

[10] D. Ruelle and F. Takens, "On the nature of turbulence," *Les rencontres physiciens-mathématiciens de Strasbourg-RCP25*, vol. 12, pp. 1–44, 1971.

[11] J. Gleick, *Chaos: The amazing science of the unpredictable*. Vintage Publishing, 1998.

[12] M. J. Feigenbaum, "Quantitative universality for a class of nonlinear transformations," *Journal of statistical physics*, vol. 19, no. 1, pp. 25–52, 1978.

[13] L. Boccato, *Novas propostas e aplicações de redes neurais com estados de eco*. Tese (doutorado), Universidade Estadual de Campinas, Faculdade de Engenharia Elétrica e de Computação, Campinas, SP, 2013.

[14] I. B. Yildiz, H. Jaeger, and S. J. Kiebel, "Re-visiting the echo state property," *Neural networks*, vol. 35, pp. 1–9, 2012.

[15] C. Gallicchio, A. Micheli, and L. Pedrelli, "Design of deep echo state networks," *Neural Networks*, vol. 108, pp. 33–47, 2018.

[16] P. Alexandroff, "Mémoire sur les espaces topologiques compacts," *Verh. Konink. Acad. Wetensch. Amsterdam*, vol. 14, pp. 1–96, 1929.

[17] H. Jaeger, "The "echo state" approach to analysing and training recurrent neural networks-with an erratum note," *Bonn, Germany: German National Research Center for Information Technology GMD Technical Report*, vol. 148, no. 34, p. 13, 2001.

[18] H. Jaeger, "Short term memory in echo state networks. gmd-report 152," in *GMD-German National Research Institute for Computer Science (2002), http://www. faculty. jacobs-university. de/hjaeger/pubs/STMEchoStatesTechRep. pdf*, Citeseer, 2002.

[19] K. Doya *et al.*, "Bifurcations in the learning of recurrent neural networks 3," *learning (RTRL)*, vol. 3, p. 17, 1992.

[20] H. Jaeger and H. Haas, "Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless communication," *science*, vol. 304, no. 5667, pp. 78–80, 2004.